



COVER SHEET

This is a version of article published as:

van Prooijen, Marvin (2006) Metadata: Development, deployment and diffusion. In Edwards, Sylvia, Eds. *Proceedings of the 1st Annual Web Content Management Symposium: C3 Create Clarify Connect*, Queensland University of Technology, Brisbane.

Copyright 2006 (please consult author)

Accessed from <http://eprints.qut.edu.au>

METADATA: DEVELOPMENT, DEPLOYMENT AND DIFFUSION

ABSTRACT

This paper illustrates five standard developments. Ways metadata may be situated relative to the user and resource are related to the core activities standard practices are intended to support. These are analogous with traditional methods of information organisation in Libraries and archiving. The examples are AGLS, METS, OAI, ONIX and EAD.

INTRODUCTION

The purpose of this paper is to illustrate metadata standards developments for managing information resources on the web. Focus will be on rich information sources such as document type objects, and descriptive metadata in the broadest sense. Rather than dealing with metadata that describes the qualities intrinsic to resource simply, or that provides for transfer of content only, developments that address the broader context of use of a resource will be of value in the future. Standards are not ready-made. They consist of a combination of practices and information technologies developing within the context of diffusion for specific domains. As will be shown, this is a general lesson from the attempt and failure to deploy basic Dublin Core elements for the broader Web domain.

Five examples of standard developments are illustrated. These are placed in a broad business context, and related to the core activities they are intended to support. At the same time, the examples reflect the variety of ways that metadata may be deployed or situated with reference to the resource and user. The aim is to show the value of metadata as a shareable asset for managing digital information collections and the multiple uses of resources. By analogy, the developing ways of deployment may be familiar to readers familiar with the history of cataloguing rules and technologies.

The paper is structured into three main sections. A background section defines the nature of metadata and outlines the ways it may be deployed. It further gives an historical description of the inception of standards with the Dublin Core Initiative, and discusses the limits of deployment for the broader Web domain. The second section illustrates standard developments in more limited domains. The examples are: the Australian Government Locator Service (AGLS); the Metadata Encoding and Transfer Standard for Libraries (METS); the Open Archives Metadata Harvesting Protocol (OAI-MHP); ONIX for publishing, and Encoded Archival Description (EAD). Within each of the domain contexts these standards are being developed, metadata is providing support and solutions for core business activities. The final section summarises the given standard developments in terms of future potential.

METADATA: A BACKGROUND SUMMARY

Definition and purpose of metadata

As a concept, Metadata is best unpacked within the context that use may be made of resources. According to the National Information Standards Organisation, "Metadata is structured information, that describes, explains, locates, or otherwise makes it easier to

retrieve, use or manage an information resource” (NISO, 2004 p. 2). As such, metadata is a tool for managing lifecycles of information resources, a method of information organisation, and an aid to access. Resources may be content modules and information bearing objects, but also collections and services to them. Metadata holds the attributes of information resources necessary for their use and management.

Deployment of metadata

There are a number of different ways that metadata may be deployed or situated relative to the user and resource. Metadata may be ‘in’, ‘between’, or ‘out’ of the information resource. It can be embedded with the content, attached to containers of content, or be autonomous as part of a collection of records. For example:

- It may be *embedded* with the content; using meta-tag elements at the back of a web page. This is analogous to publication details carried within books.
- It may be part of a *separate document*; carried in files linked to the resource it describes. Much like electronic catalogue references moving around library systems or the erstwhile catalogue card.
- Metadata Records may be stored in a database *repository*. These may be open to user access and provide links to the resource, thereby acting as a catalogue or index.
- Metadata elements may manage a *feed* for selective download prior to use of resources.

Since metadata can facilitate the selective transfer of digital content, it is able to support points in business processes and task activities. It is incumbent on any of the five developing standard therefore, to be able to address the full *context of use* of information resources. This will be seen when illustrating the five examples of standards reflecting various way of deployment in further detail.

Uses of metadata

In all cases, metadata becomes useful because it can act as a surrogate to the resources, *prior* to having full detailed knowledge of them. Metadata describes, classifies or annotates those attributes of the information object that allow for the management and use of items and collections. Records may be shared for multi purpose re-use without prior full access (NAA, 2002). This facilitates content management for:

- Identification.
- Preservation within a context of use.
- Administration of access, ownership and rules of use.
- Discovery, retrieval, and delivery.
- Web content interoperability (Duff & McKemmish, 2000).

Functional autonomy allows metadata to become a value-adding asset. This is because metadata may thereby be more easily shared and developed for value-adding descriptions considering a fuller context of user needs. For example, embedded meta-tags invisible to users and spread across individual resources, are less useful than, or without, a separate tool such as METS which allows users to apply centralised metadata

in management. Furthermore, by functional autonomy, metadata itself becomes a manageable resource asset in incremental repositories.

The Dublin Core Metadata Initiative and the Web

A historical description of standards development and the failed deployment of Dublin Core elements on the Web, provides some background and tells us about some of the things that may be required for successful implementations. Dublin Core (DC) is both an international consensus building initiative and a set of elements for deployment. It remains the basis guiding many current metadata implementations.

Initiatives in formal metadata standards development for Web content started with the Dublin Core Metadata workshops in 1995. Information professionals involved with metadata from multiple disciplines and sectors combined their ideas (Dempsey & Weibel, 1996). Experts were included from librarianship, computer science, text encoding, the museum community, publishing, and other related fields. The Initiative aims to build the broadest possible consensus for input between stakeholders to guide basic definitions of elements for up-take and diffusion.

Initially a core set of 13 descriptive metadata elements was put forward which was later extended to 15 (Daniel, Miller & Weibel, 1995). The main principle guiding design was potential for broad interoperability, as well as flexibility for further extension through implementations in practice by various constituencies of use. As a synthesis, its elements are a minimum working set for resource discovery; somewhere between the highly developed MARC standard, and a full-text search on a search engine.

One of the larger aims of the Dublin Core Metadata Initiative was to bring some structure to the World Wide Web. Though never intended for the broader Web domain alone, one hope that was important in driving collaborative development was that future use of metadata would improve precision in discovery and retrieval of Web content (Daniel, Miller & Weibel, 1995). Perhaps the problem of 'high recall' and 'low precision' of results received in search engines might be alleviated. In this way metadata could be a tool to combat the experience of information overload.

In practice the flexible nature of DC was unsuitable for use on the open Web. There was limited deployment by Web site publishers, little use of meta-tags by search engines, and negligible effects on resources discovered. Some Search Engines experimented with 'description' and 'keyword' tags. Recovered hits had a 30% adoption of these tags embedded, but comparative improvement relative to other discovery tools for discovery remained unimpressive. Lack of a controlled vocabulary meant no standard of precision could be achieved. Furthermore, publishers competing for visibility in results rankings took advantage of the ability of engines to return hits by number of entered items and alternate spellings (Alimohammadi, 2005). For example, spamming of the HTML keywords meta-tag hidden to display could sell cars with 'sex'.

Optimism regarding the deployment of metadata for the broader web community waned by the end of the nineties. From the mid nineties the journal literature was full of work on developing metadata standards for use on the Web. By 2000 there was virtually none. Search Engines dropped the 'keyword' tag and started penalising those who tried to trick the system (Alimohammadi, 2005). Ex-idealists wrote articles with such titles

as: “Metacrap: Putting the torch to seven straw-men of the meta-utopia” (Doctorow, 2001). At the same time, there were those determined to move on and leave the past behind, as suggested by a title like: “MARC must die” (Tennant, 2004). Stakeholders from business sectors and collaborators in standards initiatives regrouped and refocused their efforts.

There are a number of qualities that made the initial consensus less useful as a working standard. DC put forward an extensible set for broad semantic interoperability and easy embedment within HTML documents, without addressing further requirements for implementation. Lessons were learnt by the failure of effective uptake on the Web.

Some lessons towards improvement relate to the element set itself. The original set was limited to describing qualities intrinsic to the resource. It was poor in structural and administrative metadata that could provide for access into layers of content or provide for elements addressing the broader context of use of resources. It could not effectively map to levels of content within a compound resource, nor be used to manage security and access to the resource by defining ownership. The set needed further element definitions.

Other lessons relate to technological support and governance. The Web is a globalised, privatised wild-west, not a social democracy under the law. Since the adoption of any element by publishers of content on the Web is optional, the effective meaning of standard compliancy can be very low (UKOLN, 1997). In a large environment lacking governance cohesion or policy guidance, and without defined vocabularies for value input, no benchmark practices could be established. It becomes clear that for integrity of embedded metadata, it is desirable authoring be automated and guided with tools; preferably at pre-publishing or during various stages of lifecycle handling prior to use. Metadata requires benchmarks practices within controllable environments.

Metadata embedded within HTML is insufficient for improved resource discovery precision. Web crawlers are unlikely to look deep into any hierarchy of how metadata hooks and tags may be organised in websites. DC now refers to such use of the basic element set without addressing further factors of implementation as "metadata pidgin for digital tourists" (Hillman, 2005). Metadata enabled searching within smaller domains or using other methods of deployment is a different matter.

STANDARD DEVELOPMENTS: SPECIFIC NEEDS AND SOLUTIONS

Metadata for constituencies of use

Smaller domains with well developed systems are more suitable environments for the development of metadata standards and practices. At the domain level of companies, governments, institutions and sectors of industry, the development of metadata for the management of content can begin to match demands. A combination of drivers to innovation becomes effective at this level:

- The need for integration with core task activities.
- The need to manage information overload and explosion of digital content.
- The need to engage with cycles of evolving information technology capability.
- The need for management control over business processes.

For many organisations, metadata technologies and practices have become a necessity. When web publication was still in its early stages, a single webmaster would manage the content lifecycle. Today, as applications move to the web and content is exploding, the aim is to decentralise stewardship over digital resources and distribute authoring of content. This requires tools supporting content authoring and information resource organisation across multiple systems layers. Consequently, organisations require standard metadata for coordinating management of digital resources and to control business processes.

In the context of the domains of government, Libraries, academia, publishing and archives, different ways of deploying metadata are providing solutions to specific business problems and needs. In each of the standard developments, metadata is giving support to core business activities within domains.

The Australian Government and AGLS diffusion

In the case of government, metadata needs to support the management of records and services for public access and preservation. The Australian Government Locator Service (AGLS) is an implementation adapting Dublin Core to purpose. It employs the basic element set and remains wholly interoperable with Dublin Core. AGLS attempts to maintain the ideal of the broadest possible diffusion of a minimum interoperable standard of metadata to international specifications. However, there are a number of effective improvements aiming to achieve a minimum working standard for practice and precision (NAA, 2002).

Compared to DC, AGLS is extended, qualified and refined, enabling it to describe more categories and allow richer description of resources to higher degrees of precision. An element set of 19 is employed, with extra elements for administrative metadata. These are able to define, describe and manage the broader context of use for resources. For example: the location of ownership, or the authority of a document. Security, rights and access management, as well as version controls are thereby improved. In AGLS, elements can be amended with qualifiers for developing semantic precision. For example: rather than entering a value to a “date” element simply, one could say, “date modified”, “issued” or “authorised”. Also, value input to elements may be refined through the use of controlled vocabularies, formal thesauri and international standard schemes. Furthermore, element application to resources and value input may be automated or guided with forms based reference at or prior to authoring, as well as consequently for other points of use (NAA, 2002).

The implementation of AGLS technology and benchmarking of practices is supported by persistent governance. Its use is mandatory in government agencies as well as open for cross-domain adoption. Five metadata elements are mandatory and must be present for compliance with this standard within government agencies. Two are conditional and mandatory depending on context of use. Further benchmarking to raise standards may be implemented for specific agencies. Coverage includes web resources, services and people. Public rights to access are safeguarded in the shape of legal requirement for publishing and archiving both content and metadata. Implementation, maintenance, and metadata storage is coordinated by governance, supported by law and driven by policy, with good potential for diffusion on a national scale across domains.

Libraries and the METS package

Standards facilitate interoperability. The broadest possible diffusion is therefore the aim. However, interoperability between individual versions of standards is a major concern. Libraries manage multiple collections, with items of variable format, and are required therefore, to work with a number of standard technologies. One approach to dealing with achieving granularity of access across different standards is the Metadata Encoding and Transmission Standard. A number of research libraries are collaborating towards this in a process managed by the Library of Congress (Cundiff, 2004). What they seem to be developing is an electronic catalogue card.

METS is an XML schema document, that carries files of structural, descriptive, administrative, technical object metadata elements. The metadata, as well as any data itself, may be wrapped as part of the METS package or be referenced and located externally. A descriptive file may hold multiple record types, such as MARC or DC, for cross-referencing (Tennant, 2004). METS provides for structural elements giving access to content within resources. The structural file container defines the hierarchy of a digital library object, and enables reference to digital content files, such as pages of contents and chapters, as well as links to image, audio or video files (Cundiff, 2004). The value of METS lies in the fact that it is a tool that can be shared in use, for managing and providing access to items and collections.

Academia, research and the OAI repository

Within the academic domain, there is a strong demand for exchange of information on current research. The Open Archives Initiative supports this demand by providing a protocol that is middleware neutral. A great deal of metadata in computing is used only in embedded form for identification of data to facilitate its transfer. This can be like having publishing information within a book, without having a catalogue that is open to use for identification of its location and without the ability to share that catalogue as a whole with other organisations. The OAI allows metadata to be put to work as a separate asset.

The protocol asks participating organisations to translate local forms of object metadata into a core interoperable set for harvesting (Eden, 2002). The aim of OAI is to build federated repositories of resource identification information that are accessible to end users. Traditionally individual records have been passed between systems using messaging protocols. The harvesting protocol was originally conceived as a way to share access to web-accessible pre-print archives between research communities, thereby avoiding the costs, delays and inefficiencies of publishing infrastructure and systems middleware.

However, the concepts may apply equally to multiple formats and for different communities (Guy & Hunter, 2004). Repositories are being built beyond the academic sector. For example, Google Scholar is harvesting metadata exposed by the National Library of Australia (NLA, 2006). Participating organisations may provide linkage from the metadata to the full content of the resources (Guy & Hunter, 2004). Therefore, the Open Archives Initiative provides a way to build new versions of shared catalogues and digital collections for web delivery.

The Publishing and audio-visual sectors and the ONIX feed.

In the audio-visual and publishing sectors, ICT convergence is driving demand for new products globally. The speed of innovation and the relative inexpensiveness of publishing product having metadata lacking integrity, mean a multitude of standards. "Format wars" necessitate overarching ways to manage new content. Consequently there is a need for establishing control over production and marketing.

The driving forces behind metadata development are the proliferation and overlaps of channels and formats of content delivery, and the concomitant issues of digital rights management. Metadata is important because on the one hand, bibliographic descriptions run parallel to legal definitions of rights, while on the other hand, metadata facilitates the selective through-put of digital content in business processes (UKOLN, 2001).

The instrument gaining broad acceptance is the ONIX publishing standard. While originally developed for books, it is increasingly used for any format (UKOLN, 2001). Onix metadata may be thought of as running parallel to a chain running from product inception to end user delivery. Similar to an XML Schema, it manages the feed of publishing information transfer for local download and web rendering. Over 200 elements may be used, with basic identifier and bibliographic description mandatory. Its Elements provide for structural access, format merchandising and links to media. This may include such items as author biographies, pages of contents, chapters and links to video and websites (Brand, Daley & Meyers, 2003). There is no limit in principle to the inclusion of value-adding content. What is of greatest interest is that elements may be added at points of the chain, where there are different user needs. One such location for example, may be a Library portal.

Archives and the bridge to EAD.

It is incumbent on metadata technologies to address those practices that address the core uses made of resources for consolidation. In some cases the practical core context may be very broad indeed. In archiving, legacy information organisation is closely integrated with the context of use for resources. For example, the layers of historical annotation appended with a painting or diary, are part of their very interpretability. In Archiving, resources are organised by context of provenance, original order received, and accrued description of use. There are thick interrelationships between item and collection level units which build meaning. A metadata standard therefore has to be able to absorb this.

Encoded Archival Description (EAD) is a metadata schema reflecting archival practice. It is able to generate complex relational hierarchies of finding aids (Kiesling, 2001). A further use of EAD is in the digitalisation of content for web display. Repositories may provide digests of collections for embedment in the finding aid, or link to external digital documents (Kiesling, 2001). EAD therefore allows for the building of parallel digital collections for broad accessibility.

Archival practices consolidated in EAD create some problems for resource discovery and interoperability with other standards. Element descriptions are articulated as deep nested hierarchies. Web crawlers locate finding aids but are unlikely to access to the levels of items. One method to support interoperability has been to establish crosswalks

between EAD and other schemes. Crosswalks match and pair input to elements according to one metadata standard, and translate it according to another for output (Eden, 2002).

CONCLUSION

This paper has illustrated a variety of ways metadata may be deployed. Since 1995, the principle ideals and basic elements arising from the Dublin Core Initiative have been developed further for particular constituencies of use. Metadata technologies and practices are being implemented within limited domains to provide support for various business activities.

There are a number of general factors supporting development of standard technologies for diffusion. In order for implementation of schemes and standard practices to be successful, support from, and participation with, overarching international organisations is required; Organisations such as Dublin Core and the ISO. Success requires taking cues from those organisations in positions of leadership in practice; such as the National Archives or the Online Computer Library. It *must* involve effective governance and systems and tool support within domains. Large enterprises, federated professional communities and governments are the likely organisations able to apply the resources required. More than anything, it requires the establishment of need for activities, open participation with stakeholders, and consensus building between communities of practice according to common goals.

Each of the illustrated solutions has its strengths and problems. AGLS is developing for diffusion in an environment of persistent governance, though the technological tools are yet to be fully in place. METS provides a tool that can be shared for management of collections. There are no rules yet as to what will be contained or referenced, and interoperability problems between libraries remain. However it also took some decades to nationally standardise catalogue cards. The OA protocol enables new shared collections and indexes. An ongoing issue here is the balance of rights for access and management between users, libraries, publishers or Google. Onix metadata is added along a chain from publication at different points prior to any end use. In principle, legal instruments similar to Cataloguing in Publication may provide support for a balance favourable to information management organisations in future.

Developing technologies need to address the contexts of use for resources. Metadata becomes useful because it can act as a surrogate to resources. Given this functional autonomy, description is able to address multiple purposes of use for a resource. As such, it becomes a tool for management and access to items and collections.

REFERENCES

Alimohammadi, D. (2005). Meta-tags: still a matter of opinion [Electronic version]. *The Electronic Library*, 23(6), 625-631.

Brand, A., Daley, F. & Meyers, B. (2003). *Metadata Demystified: A guide for publishers*. Hanover Pennsylvania: Sheridan press and NISO press.

- Cundiff, M. V. (2004). An introduction to the Metadata Encoding and Transmission Standard (METS) [Electronic version]. *Library Hi Tech*, 22(1), 52-65.
- Daniel, R., Miller, J. & Weibel, S. (1995). *OCLC/NCSA metadata workshop report. OCLC, March 1995*. Retrieved March 17, 2006. from http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html
- Dempsey, L. & Weibel, S.L. (1996). The Warwick Metadata Workshop: A framework for the deployment of resource description [Electronic version]. *D- Lib Magazine*, 2(6). Retrieved March 17, 2006, from <http://www.ukoln.ac.uk/dlib/dlib/july96/07weibel.html>
- Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. V. 1.3. Retrieved March 16, 2006, from <http://en.wikipedia.org/wiki/Metadata>
- Duff, W., & McKemmish, S. (2000). Metadata & ISO 9000 compliance [Electronic version]. *Information Management Journal*, 34(1), 4-13.
- Eden, B. (2002). Applications of Metadata [Electronic version]. *Library Technology Reports*, 38(5), 60.
- Guy, M. & Hunter, P. (2004). Metadata for harvesting: The Open Archives Initiative, and how to find things on the Web [Electronic version]. *The Electronic Library*, 22(2), 168-174.
- Hillmann, D. (2005). *Using Dublin Core*. Retrieved May 6, 2006. from DCMI website: <http://au.dublincore.org/documents/2005/11/07/usageguide/index.html>
- Kiesling, K. (2001). Metadata, metadata, everywhere – but where is the hook? [Electronic version]. *OCLC Systems & Services*, 17(2), 84-88.
- Kilgour, G. (1992). Entrepreneurial Leadership [Electronic version]. *Library Trends*, 40(3), 457-474.
- NAA: National Archives of Australia. (2002). *AGLS Metadata Element Set Part 2: Usage Guide: A non-technical guide to using AGLS metadata for describing resources* V. 1.3. Retrieved March 20, 2006, from http://www.naa.gov.au/recordkeeping/gov_online/agls/metadata_element_set.html
- National Library of Australia (2006). National Library of Australia Digital Object Repository. Retrieved March 20, 2006, from <http://www.nla.gov.au/digicoll/oai/index.html>
- NISO: National Information Standards Organisation. (2004). *Understanding Metadata*. Retrieved March 17 from www.niso.org/standards/resources/UnderstandingMetadata.pdf
- Tennant, R. (2004). A bibliographic metadata infrastructure for the twenty-first century [Electronic version]. *Library Hi Tech*, 22 (2), 175-181.